**Summary of specification curve analysis:**

**(1) Identify the set of theoretically justified, statistically valid and non-redundant specifications**

* Sensible tests of the research question
* Expected to be statistically valid
* Not redundant with other specifications in the set.

The set of reasonable specifications can be generated by (1) enumerating all of the data analytic decisions necessary to map the scientific hypothesis or construct of interest onto a statistical hypothesis; (2) enumerating all the reasonable alternative ways a researcher may make those decisions; and (3) generating the exhaustive combination of decisions, eliminating combinations that are invalid or redundant. If the resulting set is too large, then in the next step (estimation) one can randomly draw from them to create specification curves.

**(2) Estimate and displaying the results graphically, allowing readers to identify consequential specifications decisions**

* The approach consists of estimating all reasonable specifications, and can display the results for all (or a large random subset of) reasonable specifications.
* The specification ‘curve’ shows the estimated effect size across all specifications, sorted by magnitude, accompanied below by a ‘dashboard chart’ indicating the operationalizations behind each result.
* Visual identification by the reader of both variation in effect size across specifications and its covariation with operationalization decisions.
* Identifying which analytical decisions are the most consequential
* This increases statistical efficiency by reducing specification noise. It also makes transparent the existence of such noise, and allows readers to determine its nature (that is, which operationalization decisions are versus are not consequential). Specification curve analysis generates a *π* with fewer arbitrary inclusion decisions, and thus more closely approximates a random sample of Π.

**A formal presentation of the problem we want to solve**

Learning about y = F(x,Z) poses several practical challenges: (1) x and y are often imprecisely defined latent variables (for example, education and economic success are both imprecisely defined latent variables); (2) the set of moderators and confounders in Z are often not fully known ex ante; (3) Z also contains imprecisely defined latent variables (for example, school quality is a latent and imprecisely defined predictor of economic success); and (4) the functional form F() is not known. To study y = F(x,Z), researchers must operationalize the underlying constructs. Let’s designate the operationalization of a construct θ, with θ↔. Researchers, then, approximate y = F(x,Z) with a specification, a set of operationalizations: y↔ky=F↔kF(x↔kx;Z↔kZ), where ky, kF, kx and kZ are indices for single operationalizations of the respective constructs. For example y↔1 may operationalize ‘economic success’ with yearly salary, while y↔2 with private jet seat capacity.

For each construct there are multiple statistically valid, theoretically justified and non-redundant operationalizations. Their combination leads to what we refer to as the set of reasonable specifications, which, as discussed in the previous section, may be at least somewhat subjective.

Designating the total number of valid operationalizations for each construct with ny, nx, nZ and nF, the total number of reasonable specifications available to study y = F(x,Z) is N ≤ nx × ny × nZ × nF.

Let Π be this set of N reasonable specifications, and π be the subset of specifications reported in a paper. Thinking about π as a sample of Π makes it easier to understand the problem that specification curve analysis attempts to remedy.

By definition, any given y↔ky=F↔kF(x↔kx;Z↔kZ)y↔ky=F↔kF(x↔kx;Z↔kZ) is considered a valid proxy for y = F(x,Z) and therefore so is the full set of all such proxies: Π. A (1) large, (2) random and (3) independently drawn sample of Π would thus lead to a reasonable estimate of the model of interest: y = F(x,Z). The problem is that π, the sample of specifications reported in a paper, has none of these three properties.

The null hypothesis that the true effect of x on y is zero for all specifications is thus: H0: d(F↔kF)d(x↔kx)d(F↔kF)d(x↔kx) = 0, ∀ πk in Π, where πk indexes the valid operationalizations in Π. For example, considering the special (though quite general) case of a general additive model where F(x,Z) = fx(x) + fz(z) + fxZ(xZ), the null is H0: d(f↔xkF)d(x↔kx)=d(f↔xZkF)d(x↔kx)d(f↔xkF)d(x↔kx)=d(f↔xZkF)d(x↔kx) = 0, ∀ πk in Π and ∀ observable x.

**(3)** **Conducting joint inference across all specifications, and using an inferential specification curve.**

* The third step answers the question: considering the full set of reasonable specifications jointly, how inconsistent are the results with the null hypothesis of no effect?
* Specification curve analysis allows statistical inference that takes into account the statistical dependence across alternative specifications in π.
* Implementing the testing of this null requires a test statistic, a single scalar on which we can measure the extremity of the data.
* The first consists of obtaining the median effect estimated across all specifications, and then testing whether this median estimated effect is more extreme than would be expected if all specifications had a true effect of zero.
* The second test statistic consists of the share of specifications that obtain a statistically significant effect in the predicted direction, testing whether such share is more extreme (higher) than would be expected if all specifications had an effect of zero.
* The third test statistic is similar to the second, but rather than discretizing each *P* value into a significant versus non-significant dichotomous variable, and counting them, it aggregates all of them in a continuous fashion, by averaging the *Z* value associated with each (for example, *Z* = 1.96 for *P* = 0.05), as in Stouffer’s method, and testing whether the average *Z* value across all specifications is more extreme than would be expected if the true effect were zero in all specifications.
* The author proposes reporting both the second and third statistics because each have its own benefit.
* Each overall P value is computed by the proportion of shuffled samples leading to a test statistic at least as extreme as in the observed sample. For P value calculations, we divide by two the proportion of shuffled samples, resulting in a test statistic of the exact same value as that in the observed data[34](https://www-nature-com.ezp-prod1.hul.harvard.edu/articles/s41562-020-0912-z#ref-CR34). When no shuffled sample is as extreme as the observed, we report P < 0.002 because our estimate is that it is less frequent than 1 out of the 500 samples we collected. However, estimates as small as that are more susceptible to random simulation error. Stouffer’s Z is computed by converting each P value to a Z-score (normal deviate) and then computing a weighted average, where the weight is 1 divided by the square root of the number of tests. The P value associated with this is also obtained via resampling, rather than from the normal distribution, to take into account the lack of independence across specifications (which is why Z = 9.22 (last row in Table [2](https://www-nature-com.ezp-prod1.hul.harvard.edu/articles/s41562-020-0912-z#Tab2)) has a non-significant P value).
* It is simple to generate such distributions by relying on resampling under-the-null. This involves modifying the observed data so that the null hypothesis is known to be true, and then drawing random samples of the modified data. The test statistic of interest is then computed on each of those samples. The resulting distribution is the estimated distribution of the test statistic under the null.
* The implementation of under-the-null resampling is more intuitive for experiments than for non-experiments, where covariates are possibly correlated with the predictor of interest. The two examples in this paper involve experiments and we thus explain resampling for experiments in this section. Resampling for observational data is discussed in more detail in in Supplementary Notes [1](https://www-nature-com.ezp-prod1.hul.harvard.edu/articles/s41562-020-0912-z#MOESM1)–[5](https://www-nature-com.ezp-prod1.hul.harvard.edu/articles/s41562-020-0912-z#MOESM1).

**Experiment data joint inference:**

Resampling experimental data under the null is simple and intuitive, as it involves shuffling the column(s) with the randomly assigned variable(s)[29](https://www-nature-com.ezp-prod1.hul.harvard.edu/articles/s41562-020-0912-z#ref-CR29),[30](https://www-nature-com.ezp-prod1.hul.harvard.edu/articles/s41562-020-0912-z#ref-CR30),[31](https://www-nature-com.ezp-prod1.hul.harvard.edu/articles/s41562-020-0912-z#ref-CR31),[32](https://www-nature-com.ezp-prod1.hul.harvard.edu/articles/s41562-020-0912-z#ref-CR32). In the case of the hurricanes paper, one shuffles the hurricane’s name. The shuffled datasets maintain all the other features of the original (for example, collinearity, time trends, skewness and so on) except that we now know there is no link between (shuffled) names and fatalities; the null is true by construction. For each shuffled dataset we estimate all 1,728 specifications. Repeating this exercise many times gives us the distribution of specification curves under the null. The only assumption behind this test is exchangeability[31](https://www-nature-com.ezp-prod1.hul.harvard.edu/articles/s41562-020-0912-z#ref-CR31),[32](https://www-nature-com.ezp-prod1.hul.harvard.edu/articles/s41562-020-0912-z#ref-CR32), that any hurricane could have received any name. The resulting *P* values are hence ‘exact’, not dependent on distributional assumptions.

**Non-Experiment data joint inference:**

To force the null on non-experimental data, we propose the following procedure, which is nearly equivalent to that of Flachaire33. For each specification one first estimates the model with the observed data—say, estimating the parameters a, b and c in y = a + bx + cz + e. Then one forces the null on the data by creating a new dependent variable, y\*, that subtracts the estimated effect of x on y—that is, y\* = y – b^x, where b^ is the sample estimate of b. With y\* we now have a model where the null is true—that is, we have y\* = a + b\*x + cz + e, where we know that b\* = 0.

To generate a distribution of expected results, the sampling distribution of b^ under the null, one samples with replacement rows of data by using y\* rather than y as the dependent variable. Each resample has the same sample size as the original. The resulting distribution of b^ across the resamples is used to assess the extremity of the observed b^ if the null were true. Applying this approach to specification curve analysis leads to the following six steps:

(1)

Estimate all K specifications with the observed data, y↔ky=F↔kF(x↔kx;Z↔kZ). These will result in K different point estimates: b^k. with k = 1…K. Note that y↔ky may be the same for more than one specification, even for all K of them, if the operationalization of the dependent variable is not varied across specifications.

(2)

Generate K different dependent variables under the null, yk\* = yk −bˆk×xk. Even if there are fewer than K different yk, there will be K different y\*k because bˆk is different across specifications and thus so is y\*k. So now every row of data has the x values and K different y\* values.

(3)

Draw at random, and with replacement, N rows from this matrix, using the same drawn rows of data for all K specifications.

(4)

Estimate the K specifications on the drawn data.

(5)

Repeat steps 3 and 4 a large number of times (for example, 500 or 1,000).

(6)

For each bootstrapped sample we now have K estimates, one for each specification. Compute what percentage of the resampled specification curves (for example, of the 500 resamples) exhibits an overall test statistic (for example, median effect size) that is at least as extreme as that observed in the real data.